Advanced Data Cleaning with Fuzzy String Matching for Data Deduplication

1. Task Description

The objective of this task is to perform advanced data cleaning techniques on a dataset, specifically focusing on **fuzzy string matching for data deduplication**. This process is essential in cleaning datasets that may contain duplicate or similar entries with slight variations in text. The dataset contains a collection of text-based records that may contain duplicates. The aim is to identify and remove such duplicates based on approximate matches of text entries.

2. Screenshot of Output

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	34.5	0	0	7.8292	Q
1	1	3	female	47.0	1	0	7.0000	S
2	0	2	male	62.0	0	0	9.6875	Q
3	0	3	male	27.0	0	0	8.6625	S
4	1	3	female	22.0	1	1	12.2875	S
...	...	...	...	...	...	...	...	...
413	0	3	male	27.0	0	0	8.0500	S
414	1	1	female	39.0	0	0	108.9000	C
415	0	3	male	38.5	0	0	7.2500	S
416	0	3	male	27.0	0	0	8.0500	S
417	0	3	male	27.0	1	1	22.3583	C

418 rows × 8 columns

3. Algorithm Used in Task

- **Libraries and Algorithms:**
  - **Fuzzy String Matching:**
    - To perform data deduplication, the **fuzzywuzzy** library was utilized, which allows for approximate string matching.
    - This technique is valuable when there are slight differences between duplicate strings, such as variations in spacing, spelling, or punctuation.

4. Add Report in Your Task Zip File

The task report has been added to the zip file. This includes:

1. The Python script (Data Science & Machine Learning_Task_3.ipynb).
2. A text version of this report (Task_3_Report.pdf).

Prepared by
Karan Gohil